Discovering the Unknown Suggestion: a Short Review on Explainability for Recommender Systems

Luca Marconi^{1,2}, Ricardo A. Matamoros A.^{1,2} and Francesco Epifania²

¹Department of Informatics, Systems and Communication, University of Milano - Bicocca, Milan, Italy ²Social Things SRL, Italy

Abstract

Artificial Intelligence and in particular machine learning and deep learning models are normally considered to be fast and high performing, but in general there is a lack of transparency and interpretability. The issues related to explainability and its consequences are becoming more and more relevant in the whole broad scenario of Artificial Intelligence. To address this issue, explainable AI emerged, as a set of Artificial Intelligence techniques able to make their own decision more transparent and interpretable, so as to let users understand the specific reasons why the system provided its outcome, decision, or, in the case of recommender systems, its suggestions. Explainable Artificial Intelligence is deeply needed in heterogeneous domains and contexts, as the need for transparency, interpretability and even accountability of the Artificial Intelligence-based systems is a big necessity, as confirmed by the recent right to explanation in the 2018 General Data Protection Regulation by the European Union. Due to the diffusion of recommender systems in many applicative domains and situations in everyday life and business fields, there is an emerging necessity for systems not only able to provide human decision-makers with suggestions and ease the decision-making processes in organizations, but also to give the right motivations of their recommendations. This paper summarizes the results of the study of the state of the art for Explainable Artificial Intelligence for Recommender Systems. We will follow the main reviews in literature to present the main work, kinds of explanainable recommendations and methods.

Keywords

Machine Learning, Recommender Systems, Artificial Intelligence, eXplainable Artificial Intelligence, eXplainable Recommender Systems

1. Introduction

Nowadays, Artificial Intelligence (AI) is becoming more and more important in our professional and personal life. According to the International Data Corporation (IDC) the global investment on AI will reach almost 118 billion U.S. dollars in 2022 and even surpass 300 billion U.S. dollars by 2026 [1]. Moreover, the statistics portal Statista forecasts that revenues from the AI market worldwide will grow from 10.1 billion U.S. dollars in 2018 to 126 billion U.S. dollars by 2026 [2]. Gartner identifies AI as a fundamental technology in most of the the Gartner Top 10 Strategic Technology Trends for 2023 [3]. In the context of the current fourth industrial revolution, overlapping waves of breakthroughs in computing, artificial intelligence, nanotechnology and

[☑] l.marconi3@campus.unimib.it (L. Marconi); r.matamorosaragon@campus.unimib.it (R. A. M. A.); r.matamorosaragon@campus.unimib.it (F. Epifania)

D 0000-0002-0877-7063 (L. Marconi); 0000-0002-1957-2530 (R. A. M. A.); 0000-0002-1957-2530 (F. Epifania)

CEUR Workshop Proceedings (CEUR-WS.org)

material science, 3D-printing, molecular biology (gene sequencing), robotics and other evolving and emergent technologies are reshaping life, business models and ecosystems, according to [4] In this scenario, AI is strongly emerging as transversal and powerful technological paradigm, due to its ability not only to deal with data and big data, but especially because it produces and manages knowledge. Andrew Ng, former chief scientist at Badu and Co-founder at Coursera, said in a keynote speech at the AI Frontiers conference in 2017 that AI is really the new electricity: a disruptive, pervasive and enabling technology, empowering technologies and processes in potentially any field or domain. AI and in particular Machine Learning (ML) and Deep Learning (DL) models are normally considered to be fast and high-performing, but in general there is a lack of transparency and interpretability [5, 6, 7]: it's hard work to get insights from their internal mechanisms when trying to understand why the system provided its outcome or decision. To address this issue, explainable AI (XAI) emerged, as a set of AI techniques able to make their own decision more transparent and interpretable, so as to let users understand the specific reasons why the system provided its outcome, decision, or, in the case of recommender systems, its suggestions [5, 6, 7]. Explainable AI is deeply needed in heterogeneous domains and contexts, as the need for transparency, interpretability and even accountability of the AI-based systems is a big necessity, as confirmed by the recent right to explanation in the 2018 General Data Protection Regulation (GDPR) by the European Union [8]. Due to the diffusion of Recommender Systems (RSs) in many applicative domains and situations in everyday life and business fields, there is an emerging necessity for systems not only able to provide human decision-makers with suggestions and ease the decision-making processes in organizations, but also to give the right motivations of their recommendations [9, 10]. A good way to classify eXplainable Recommender Systems (XRSs) was proposed by Zhang et al. in 2014 [11]: it essentially deals with two dimensions: the information source or display style of the explanations (e.g., textual sentence explanation, or visual explanation): it represents the humancomputer interaction perspective of explainable recommendation research; the model itself, representing the machine learning perspective of explainable recommendation research. XRSs can be evaluated both by qualitative, user-centered and quantitative evaluation methods. The evaluation can be either related to the performance of the system or to its explainability. In both cases, experiments can be designed wher real users are involved, or without the contribution of human users in the experimental setting. When it comes to evaluating the explainability of the RSs. methods can regard online, offline evaluation or user studies [12], while other classifications have been proposed in the literature. Overall, the evaluation of explainability suffers from a lack of a unified, precise and widely accepted formal definition of explainability, which implies the use of complementary qualitative and quantitative methodologies to completely strive to evaluate such systems.

This paper summarizes the results of the study of the state of the art for XRSs. We will follow the main reviews in literature to present the main work, kinds of explainable recommendations and methods. Thus, the aim of this paper is to provide a short and compact macro-review of the mostly diffused and used methods and systems reported in the literature. The rise, evolution, adaptation and modifications of models are definitively ongoing processes in the state-of-the-art, thus getting a comprehensive and complete classification is challenging. Given the ongoing evolution of the field, as well as the increasing number of potential applications, the aim of this paper is definitively not to provide a comprehensive and complete review of the large panorama of such discipline. Rather, we report a limited and carefully circumscribed set of fundamental concepts and methods to get a general picture for later understanding and appreciate the many potential applications and uses of for heterogeneous business and industrial domains. Given the increasing need of explainable, interpretable and thrustworthy systems in business and organizational Therefore, the proposed survey is intended to provide a general overview of the growing scenario of the XRSs, with the aim to help researchers, practitioners and decision-makers to orient themselves to exploit the many potentialities of explainability in recommender systems for business and industrial applications.

2. The context of Explainable AI

Actually, the explanation problem is definitively not new in the literature: the term started to be used in 2004 [13], though the problem itself has existed since the mid-1970s, specifically in the field of expert systems [14], with the first rise of AI in the literature. Though, a greater interest in this theme started to grow with the evolution of machine learning methodologies and techniques, particularly with the growth of its performances in the last years. In the literature, the need for explainable AI is motivated mainly by three reasons: the need for *trust*, for *interaction* and for *transparency* [7]. It's worth to notice that, consequently, explainable AI is strictly related to *responsibility* and transparency [7, 12, 6]. Consequently, explainability is definitively becoming a key conceptual elements for the present and incoming AI systems, as it is also explicitly required in the European General Data Protection Regulation (GDPR) [15], where also the key related concepts of *fairness* and transparency in automated decision-making are highlighted.

In general, XAI is strongly needed for justifying and interpreting the results, so as to ensure that they were not made erroneously [7, 6]. Moreover, the possibility to explain the results would help to improve the way the results are obtained, control the systems dynamics and facilitate new ways to gain knowledge [7].

In a broader perspective, the diffusion of XAI methods and techniques is a crucial step in the current and future evolution of AI systems. Such methods can significantly be grouped into the so-called *third wave* of AI, as defined by DARPA. Thus, XAI strives to realize the big challenge of contextual adaptation, i.e. the construction of progressively explanatory methods for classes of real-world phenomena. The further steps in the design and development of such new and empowered AI systems is the ability to foster continuous learning by the inclusion of synergetic learning techniques, as well as the progressive empowerment of the interaction with human decision-makers [7]. Eventually, the last mile of this ambitious evolution is the quest for reaching or emulating the human intelligence [7].

In the literature, there are different ways to classify the XAI models: among them, there are classifications distinguishing algorithms for their *global* or *local* interpretability, and classifications taking into considerations the differences between *model-specific* or *model-agnostic* methods, thus related with the possibility to apply explainable techniques only to specific models or not. We present the main useful concepts for our work, as well as the main classification reported in the literature, in the case of XRS.

3. Explainable Recommendations

In this context, explainable AI in the field of RSs is aimed at providing intuitive explanations for the suggestions and recommendations provided by the algorithms [12, 19]. Basically they try to address the problem of why certain recommendations are suggested by the models. As they are part of the big world of the XAI, explainable recommendations can either be *model-intrinsic* or *model-agnostic*: in the former case, the output model is intrinsically interpretable, meaning that the decision mechanism is completely transparent providing explainability; in the latter case, instead, the output model provides the so-called *post-hoc* explanations, without any modification of the model itself. It is interesting this two approaches can be conceptually linked to a cognitive psychological root [12]: in this perspective, the model-intrinsic models would be similar to the human minds rational decisions taken after some reasoning process, while the model-agnostic ones would somehow resemble the intuitive ways of deciding, followed by some search of the explanations.

In other words, as in the general case of XAI, XRSs, based on explainability-aware ML techniques, can generally be categorized into two main groups [18]:

- 1. Systems providing an explanation of their predictions in a way that is interpretable by the user. These types of methods usually only justify their output by the means of an added explanations, but without providing an in depth understanding of the underlying algorithm. This is typical in the case of post-hoc explanations.
- 2. Explainable systems directly incorporating interpretable models in the construction of the automated systems. Model intrinsic and, specifically, white-box models, such as DTs, can be categorized in this group.

XRSs started formally to be defined, conceived and used in recent years. The term explainable recommendation was formally introduced by Zhang et al. in 2014 [11], but there were earlier works in personalized recommendation research. An extensive review of the first historical stages of explainable recommendation and how it was focused especially on collaborative filtering methods in RSs is in Zhang et al., 2018 [12].

4. Classification of Explainable Recommender Systems

A good way to classify XRS was proposed by Zhang et al. in 2018 [12]: it essentially deals with two dimensions:

- 1. the *information source* or *display style* of the explanations (e.g., textual sentence explanation, or visual explanation): it represents the human-computer interaction perspective of explainable recommendation research;
- 2. the model itself, representing the machine learning (ML) perspective of explainable recommendation research.

A somehow generalized taxonomy, focused on the specific classification of interpretability methods, is provided in the review by Linardatos in 2021 [66], which depicts and highlights the major concepts and dimensions involved in the analysis of interpretable models. It proves to be useful to get a complete picture of the most significant conceptual perspectives involved.

4.1. Information Source for Explanations

The first dimension of this classification model is the *information source* for explanations, also called *display style*: namely, explanations are pieces of information related to the recommendations given by the algorithm. Recommendations can come from different information sources and can be displayed in several ways: some examples include textual sentences, word clouds or visual explanations. In the following paragraphs we provide a short summary of the different types of recommendation explanations and we give some examples of relevant related work.

Explanations based on Relevant Users or Items

This comes from the first stages of recommendation explanation research. User-based explanations are especially used by collaborative filtering RSs, thus when the recommendation is based on the ratings or interests of "similar" users. [12] reports the example of Herlocker et al. [20], comparing the effectiveness of different display styles for explanations in user-based collaborative filtering. Instead, for item-based explanations, the measure of similarity comes from the user's past liked items. Zhang and Chen [12] argue that relevant-item explanations are more intuitive for users than user-based explanations due to the familiarity of the user with the items more than with other potential users: nevertheless, this problem could be solved by another kind of explanations, the so-called *social explanation*.

Feature-based Explanations

This kind of explanations are especially related to content-based recommendation methods [12]. CB-RSs elaborate suggestions according to a specific match between users' proles and content features of candidate items. In this case it is more intuitive to base the recommendations on the specific features of the items, and then to display them in the best explanation style: for example in Vig et al. [21] the recommendations are provided adopting movie tags as features.

Textual Sentence Explanations

This kind of explanations is very useful for getting relevant benefits from user-generated content, such as e-commerce reviews and social media posts [12]. Sentences could come from pre-defined templates or be directly generated based on natural language generation models. Zhang and Chen [12] classify such approaches between *aspect-level* and *sentence-level* approaches, based on the display style of the explanations. It is worth to notice some sort of similarity between aspect-level textual explanations and feature-based explanation: though, in

the former case the aspects addressed are usually not directly available in an item or user prole. In fact, they come from textual information usually related to and users opinions or textual feedback about specific items. This is what happens in [11], where explanations are presented as aspect-opinion wordclouds based on large-scale user reviews.

Visual Explanations

They help users to get precise and intuitive suggestions. Visually explainable recommendation are still a relatively new topic in research, thus the integration of visual information and images into recommender systems is far from being optimized in terms of both explainability and performance [12]. For example, in [16] visually explainable recommendation are based on personalized region-of-interest high-lights.

Social Explanations

The involvement of friends in the recommendation process implies a higher level of personalization into the suggestions themselves, while solving the typical trustworthiness and privacy problems of relevant-user explanations. Examples include the studies and applications in music [17], and in product recommendations [22].

4.2. Explainable Recommendation Models

The second dimension of the classification model proposed by Zhang and Chen [12] regards the specific models used for producing the explanations: namely, explanations given by different types of algorithms. As always, explainable recommendations can either be *model-intrinsic* or *model-agnostic*. In the following paragraphs we provide a short summary of the major types of explainable recommendation models and we give some examples of relevant related work.

Factorization Models

Latent Factor Models based on Matrix Factorization is a classical ML model for recommender systems [26]. It learns latent factors to predict the missing ratings in a user-item rating matrix. Factorization models for explainable recommendations have been proposed in order to explain the specific latent factor acting user decisions. As an example, Explicit Factor Models [11] links each latent dimension of matrix factorization with an explicit feature among the users favorite ones. Thus, it can provide explicit recommendations based on the features. Instead, other studies [27], focus on model-based approaches to generate relevant-user or relevant-item explanations based on the user-item rating matrix.

Topic Modeling

This kind of explainable recommendations is still based on text information. Topic modeling

refers to a general methodology to classify semantics in documents according to topics clusters. Explanations are generally displayed in the form of topical word clouds. McAuley and Leskovec [28] proposed to use a model based on latent factor analysis to understand hidden topics learned from reviews. Other studies [29] focused on other probabilistic graphic models. Wu and Ester [29] created an hybrid model based on both collaborative filtering and aspect-based opinion mining. The algorithm analyses users preferences on item aspects according to reviews and then predicts the users ratings on different ones.

Graph-based Models

Graphs help to define relevant relations among information, so they can be specifically useful to represent user-user or user-item relationships, especially in social recommendation scenarios. For example, Park et al. [22] use a graph-based explainable recommendation algorithm for providing interpretable suggestions thanks to rating and similar users. Other authors exploited other kinds of graphs: in He et al. [30] a tripartite graph structure allows to model user-item-aspect relations where an aspect is an item feature generally taken from user reviews. These relations are constructed for the possible recommendations and then aspects are ranked and explanations are given to the top-ranked aspects matching the target user and the recommended item. Heckel et al. [31] instead created explainable recommendations thanks to over-lapping co-clustering based on user-item bipartite graph [12]: this approach allows to exploit both clusters of similar users and of items with similar properties.

Deep Learning

Given the higher and higher importance of deep learning techniques, there are many studies and experiments to adopt a huge variety of them in the explainable recommendations scenario. In Seo et al. [32] user preferences and item properties are represented through convolutional neural networks upon review text, so as to attribute specific weights to words in the text and highlight the relevant ones to provide explainable recommendations. Among the other various typologies of neural networks used, it is worth to cite the work by Chen et al. [33], where explainable sequential recommendation are extracted due to memory networks: they have memory over previous items chosen, so each item in the users interaction history is in a memory slot and predictions of the new behaviors can be made and explained subsequently, so as to directly show the way the users previous choices influenced new predictions. That implies the possibility to set dynamic explainable recommendations. Another interesting approach comes from capsule networks, namely neural networks empowered with capsule structures to manage hierarchies. Li et al. [34] use capsule networks to model item aspects and users viewpoints as logic units, so as to get the users' rating behaviors. Then, the algorithm, for each user-item pair, extracts the informative logic units from the reviews so as to infer their corresponding sentiments.

Knowledge Graph-based

As one of the classical ways to manage knowledge, knowledge graphs can be used for providing better explanations for the recommended items thanks to their information about users and items. Catherine et al. [35] proposed a method to provide explanations and recommendation after producing a rank of the items thanks to information found in knowledge graphs. Instead, Ai et al. [36] constructed a user-item knowledge graph, so as to get recommendations for a user as the most similar item under the "purchase" relation. In this way, they can establish a series of relations between users and items to orient and explain recommendations.

Data Mining

Among the various possibilities and techniques, Zhang and Chen [12] report that the most frequently used one is association rule mining. As an example, Davidson et al. [23] introduced the YouTube video recommendation system, adopting association rule mining to create associations between couples of videos co-watched within the same session. Then, explanations are given considering the seed video and the the association rules themselves. The approach for transparent, scrutable, and explainable recommendations suggested by Balog et al. [25] is particularly interesting: given a set of tags or keywords characterizing user preferences, they aimed at inferring preferences and recommendations by aggregating over items associated with a tag. Consequently, item recommendations can be both transparent and explainable. They chose to provide recommendations through sentence-level textual explanations, allowing users to provide feedback on clear and scrutable suggestions. It is worth to mention that this approach is a framework, which can be generalized to different machine learning models.

Model Agnostic and Post Hoc

These approaches are typically used when it is to difficult to include the explanability in the recommendation model itself. Then, after the recommendations have been provided, an explanation model generates the explanations according to the previously created recommendations. As an example exploiting a data mining technique (thus related to the previous paragraph), Peake and WanH [37] proposed an association rule mining approach. The method considers the users' transaction history to explain the recommendation: namely, the association rules help to associate the recommendations themselves with the users' previous choices, thus providing explanations to the recommendations.

Overall, the literature makes a clear distinction among models that are interpretable by design, and those that can be explained by means of external XAI techniques. This duality could also be regarded as the difference between interpretable models and model interpretability techniques; a more widely accepted classification is that of transparent models and post-hoc explainability.

In particular, local interpretation methods explain predictions individually from each other. Among these [38] we have:

1. Individual conditional expectation (ICE) [39] curves underlie partial dependence plots (PDPs) and describe how the change in a feature affects the change in the prediction.

- 2. Local surrogate models, as the Local Interpretable Model-agnostic Explanations model (LIME) [40] explain a prediction by replacing the complex model with an interpretable local surrogate.
- 3. Scoped rules (anchors) [41] are rules that describe which feature values allow the prediction to be fixed.
- 4. Counterfactual explanations [42, 43] explain a prediction by examining which features should be changed to achieve the desired prediction.
- 5. Shapley values [44] are an attribution method that assigns prediction equally to individual features.
- 6. SHAP (SHapley Additive exPlanations) [45] is another computation method for Shapley values, but unlike these it proposes global interpretation methods based on combining Shapley values across data.

4.3. Intrinsic, interpretable, white-box models

Hereinafter, we recall the main interpretable models. We focus on such macro-category of models due to the fact that the chosen approach for our XRS, DT models, is actually interpretable. Therefore, we synthetically show the main characteristics of these models, as well as their main advantages and disadvantages. Finally, we sketch the main motivations that lead us to orient ourselves towards a decision-tree approach.

Linear regression

A linear regression model predicts the target as a weighted sum of the feature inputs [38]. Linear regression are particularly useful and significant in practice for their linearity. They have long been used by statisticians, computer scientists, mathematicians and practitioners in general [38]. They are usually exploited to model the dependence of a regression target y on some features x, and the predicted outcome of an instance is a weighted sum of its features, where the optimal weights can be estimated by several methods.

The main advantages of such methods are its linearity and the modeling of the predictions as a weighted sum makes it transparent how predictions are produced [16]. The modeling of the predictions as a weighted sum makes guarantees transparency on how predictions are created. From the mathematic point of view, they are widely accepted and diffused methods among practitioners, and high level of collective experience and expertise is available in the scientific community [38].

Nevertheless, the are only useful for representing linear relationships, while any required nonlinearity or interaction has to be hand-crafted and explicitly provided to the model [38]. Moreover, they often have no good predictive performance, due to restricted ability to represent reality in a purely linear way [38]. Finally, there is a possible unintuitive interpretation of weights, due to the correlations and interactions with all the other involved features [38].

Logistic regression

Linear regression models the probabilities for classification problems with two outcomes. It's

an extension of the linear regression model for classification problems [38]. Therefore, it shows similar advantages and disadvantages than the linear regression models [38]. Also logistic regression has been widely used by practitioners in different domains and application fields, and it has issues with restrictive expressiveness and with dealing with interactions, as well as with limitations in predictive performance. Moreover, logistic regression can suffer from *complete separation*, namely the impossibility tobe trained in the case where there is a feature that would perfectly separate the two classes.

GLM and GAM

Generalized Linear Models (GLMs) and Generalized Additive Models (GAMs) are heterogeneous generalization models of regression, useful for modeling real-life situations. They can be applied in situations where the classical regression approaches fail or its assumptions are violated [38]. In the case of GLM, they can be applied where the input features do not follow a Gaussian distribution, which concretely happens in many cases in reality [38]. Instead, GAMs deal with the cases of nonlinearities, not tackled by the classical linear models. GAMs relax the restriction that the relationship must be a linear weighted sum, assuming that the outcome can be modeled by arbitrary functions that can be involved for each features [38]. Then, such models are generally.

In general, these models are highly flexible and useful for making predictions and inferences in many application cases and contexts. These methods are highly diffused in the scientific community and updated methods are often released allowing to make inferences for heterogeneous problems and applications [38]. Though, such models suffer from a significant reduced interpretability, as compared with the classical linear models, and they strongly rely on assumptions about the data generating process, which have to be respected for the validity of the model and its interpretation of the weights [38].

Decision-trees

Already previously introduced in this Chapter, DT models are useful for solving many of the presented issues, especially in the case of linear regression and logistic regression models, which have problems in situations where the relationship between features and outcome is nonlinear or where features interact with each other [38]. Tree-based models work through an iterative process of multiple splitting of the dataset, according to certain cutoff values in the features. Thus, they are inherently interpretable due to the tree structure itself, while they are also able to capture interactions between features in the data, as well as to effectively explain and visualize their output results. The main disadvantages are related to their inability to deal with linear relationships, as well as their lack of smoothness and unstability [38]. Moreover, their interpretability is reduced in the case of a significant increase in the tree depth [38].

Decision rules

Decision rules are probably the most interpretable models. IF-THEN statement consist of a

condition (antecedent) and a prediction and, in simple cases, they semantically resembles natural language [16]. Then, they are usually easy to interpret, expressive, robust and compact [38]. Nevertheless, in the literature they are used only for classification [38], resulting in applications for restricted classes of problems. Moreover, they necessarily require categorical features and, as in the case of decision-trees, they have issues in describing linear relationships.

5. Conclusions

In this study we strived to provide an overview of Explainable AI in the field recommender systems. We are aware that many other issues could have been addressed, specifically regarding the pros and cons of the wide set of methods in the literature, as well as the many evaluation techniques of both RS and explainability. As a general consideration to conclude our study, we definitively agree that the evolution of such systems necessarily involve a synergy between the empowerment of the models' performances and the emergent human-AI interaction perspective. We also conclude that much more work and effort should be dedicated to search and adopt a widely accepted, pre-defined and formally circumscribed definition of explainability and its related concepts. While there are several studies proposing both qualitative and quantitative definitions, it should be necessary to both choose and apply them to the field of recommender systems: this investigation could be the aim of a future study. Indeed, there are many further challenges and possible future directions to explore for this fascinating topic: among them, the issues related to the difficulties in quantiatively and formally measuring explainability, which will be a key step to reach and exploit the full potentialities of explainable and interpretable recommender systems for heterogeneous business and industrial domains. Moreover, the rise of explainable intelligent recommender systems will increasingly require to further investigate the broader impact of explainability on decision-making processes, so as to understand their full influence in organizational context and applications. We hope that our work can contribute to help researchers, scholars and practitioners to understanding the concept of explainable recommendation, the main approaches in the literature and their potentialities for business or industrial applications.

References

- [1] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).
- [2] Liu, Shanhong. "Revenues from the Artificial Intelligence (AI) Software Market Worldwide from 2018 to 2025." (2020).
- [3] CeArley, David, et al. "Top 10 strategic technology trends for 2018." The Top 10 (2016): 1-246.
- [4] Schwab, Klaus. The fourth industrial revolution. Currency, 2017.
- [5] Gunning, David. "Explainable artificial intelligence (xai)." Defense advanced research projects agency (DARPA), nd Web 2.2 (2017): 1.
- [6] Došilović, Filip Karlo, Mario Brčić, and Nikica Hlupić. "Explainable artificial intelligence: A

survey." 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO). IEEE, 2018.

- [7] Adadi, Amina, and Mohammed Berrada. "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)." IEEE access 6 (2018): 52138-52160.
- [8] Selbst, Andrew, and Julia Powles. ""Meaningful Information" and the Right to Explanation." Conference on Fairness, Accountability and Transparency. PMLR, 2018.
- [9] Vultureanu-Albişi, Alexandra, and Costin Bădică. "Recommender systems: an explainable AI perspective." 2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA). IEEE, 2021.
- [10] Sardianos, Christos, et al. "The emergence of explainability of intelligent systems: Delivering explainable and personalized recommendations for energy efficiency." International Journal of Intelligent Systems 36.2 (2021): 656-680.
- [11] Zhang, Yongfeng, et al. "Explicit factor models for explainable recommendation based on phrase-level sentiment analysis." Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. 2014.
- [12] Zhang, Yongfeng, and Xu Chen. "Explainable recommendation: A survey and new perspectives." Foundations and Trends® in Information Retrieval 14.1 (2020): 1-101.
- [13] Van Lent, Michael, William Fisher, and Michael Mancuso. "An explainable artificial intelligence system for small-unit tactical behavior." Proceedings of the national conference on artificial intelligence. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004.
- [14] Moore, Johanna D., and William R. Swartout. Explanation in expert systemss: A survey. UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST, 1988.
- [15] Goodman, Bryce, and Seth Flaxman. "European Union regulations on algorithmic decisionmaking and a "right to explanation"." AI magazine 38.3 (2017): 50-57.
- [16] Chen, Xu, et al. "Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation." Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019.
- [17] Sharma, Amit, and Dan Cosley. "Do social explanations work? Studying and modeling the effects of social explanations in recommender systems." Proceedings of the 22nd international conference on World Wide Web. 2013.
- [18] Abdollahi, Behnoush, and Olfa Nasraoui. "Transparency in fair machine learning: the case of explainable recommender systems." Human and machine learning. Springer, Cham, 2018. 21-35.
- [19] Biran, Or, and Courtenay Cotton. "Explanation and justification in machine learning: A survey." IJCAI-17 workshop on explainable AI (XAI). Vol. 8. No. 1. 2017.
- [20] Herlocker, Jonathan L., Joseph A. Konstan, and John Riedl. "Explaining collaborative filtering recommendations." Proceedings of the 2000 ACM conference on Computer supported cooperative work. 2000.
- [21] Vig, Jesse, Shilad Sen, and John Riedl. "Tagsplanations: explaining recommendations using tags." Proceedings of the 14th international conference on Intelligent user interfaces. 2009.
- [22] Park, Haekyu, et al. "Uniwalk: Explainable and accurate recommendation for rating and

network data." arXiv preprint arXiv:1710.07134 (2017).

- [23] Davidson, James, et al. "The YouTube video recommendation system." Proceedings of the fourth ACM conference on Recommender systems. 2010.
- [24] Hussein, Doaa Mohey El-Din Mohamed. "A survey on sentiment analysis challenges." Journal of King Saud University-Engineering Sciences 30.4 (2018): 330-338.
- [25] Balog, Krisztian, Filip Radlinski, and Shushan Arakelyan. "Transparent, scrutable and explainable user models for personalized recommendation." Proceedings of the 42nd international acm sigir conference on research and development in information retrieval. 2019.
- [26] Mehta, Rachana, and Keyur Rana. "A review on matrix factorization techniques in recommender systems." 2017 2nd International Conference on Communication Systems, Computing and IT Applications (CSCITA). IEEE, 2017.
- [27] Abdollahi, Behnoush, and Olfa Nasraoui. "Using explainability for constrained matrix factorization." Proceedings of the eleventh ACM conference on recommender systems. 2017.
- [28] McAuley, Julian, and Jure Leskovec. "Hidden factors and hidden topics: understanding rating dimensions with review text." Proceedings of the 7th ACM conference on Recommender systems. 2013.
- [29] Wu, Yao, and Martin Ester. "Flame: A probabilistic model combining aspect based opinion mining and collaborative filtering." Proceedings of the eighth ACM international conference on web search and data mining. 2015.
- [30] He, Xiangnan, et al. "Trirank: Review-aware explainable recommendation by modeling aspects." Proceedings of the 24th ACM international on conference on information and knowledge management. 2015.
- [31] Heckel, Reinhard, et al. "Scalable and interpretable product recommendations via overlapping co-clustering." 2017 IEEE 33rd International Conference on Data Engineering (ICDE). IEEE, 2017.
- [32] Seo, Sungyong, et al. "Interpretable convolutional neural networks with dual local and global attention for review rating prediction." Proceedings of the eleventh ACM conference on recommender systems. 2017.
- [33] Chen, Xu, et al. "Sequential recommendation with user memory networks." Proceedings of the eleventh ACM international conference on web search and data mining. 2018.
- [34] Li, Chenliang, et al. "A capsule network for recommendation and explaining what you like and dislike." Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval. 2019.
- [35] Catherine, Rose, et al. "Explainable entity-based recommendations with knowledge graphs." arXiv preprint arXiv:1707.05254 (2017).
- [36] Ai, Qingyao, et al. "Learning heterogeneous knowledge base embeddings for explainable recommendation." Algorithms 11.9 (2018): 137.
- [37] Peake, Georgina, and Jun Wang. "Explanation mining: Post hoc interpretability of latent factor models for recommendation systems." Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018.
- [38] Molnar, Christoph. Interpretable machine learning. Lulu. com, 2020.
- [39] Goldstein, Alex, et al. "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation." journal of Computational and Graphical Statis-

tics 24.1 (2015): 44-65.

- [40] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.
- [41] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Anchors: High-precision modelagnostic explanations." Proceedings of the AAAI conference on artificial intelligence. Vol. 32. No. 1. 2018.
- [42] Wachter, Sandra, Brent Mittelstadt, and Chris Russell. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR." Harv. JL & Tech. 31 (2017): 841.
- [43] Dandl, Susanne, et al. "Multi-objective counterfactual explanations." International Conference on Parallel Problem Solving from Nature. Springer, Cham, 2020.
- [44] Shapley, Lloyd S. "A value for n-person games." Classics in game theory 69 (1997).
- [45] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017).
- [46] Gorwa, Robert, Reuben Binns, and Christian Katzenbach. "Algorithmic content moderation: Technical and political challenges in the automation of platform governance." Big Data & Society 7.1 (2020): 2053951719897945.
- [47] Gillespie, Tarleton. "Content moderation, AI, and the question of scale." Big Data & Society 7.2 (2020): 2053951720943234.
- [48] Bonhard, Philip, and Martina Angela Sasse. "'Knowing me, knowing you'-Using profiles and social networking to improve recommender systems." BT Technology Journal 24.3 (2006): 84-98.
- [49] Subramanyam Kalyan, Katikapalli, Ajit Rajasekharan, and Sivanesan Sangeetha. "AMMUS: A Survey of Transformer-based Pretrained Models in Natural Language Processing." arXiv e-prints (2021): arXiv-2108.
- [50] Nuruzzaman, Mohammad, and Omar Khadeer Hussain. "A survey on chatbot implementation in customer service industry through deep neural networks." 2018 IEEE 15th International Conference on e-Business Engineering (ICEBE). IEEE, 2018.
- [51] Garg, Vishal, and Ritu Tiwari. "Hybrid massive open online course (MOOC) recommendation system using machine learning." International Conference on Recent Trends in Engineering, Science & Technology-(ICRTEST 2016). IET, 2016.
- [52] Young, Tom, et al.. "Recent trends in deep learning based natural language processing." ieee Computational intelligenCe magazine 13.3 (2018): 55-75.
- [53] Hu, Dichao. "An introductory survey on attention mechanisms in NLP problems." Proceedings of SAI Intelligent Systems Conference. Springer, Cham, 2019.
- [54] Alyafeai, Zaid, Maged Saeed AlShaibani, and Irfan Ahmad. "A survey on transfer learning in natural language processing." arXiv preprint arXiv:2007.04239 (2020).
- [55] Vedantam, Vamsi Krishna. "The Survey: Advances in Natural Language Processing using Deep Learning." Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12.4 (2021): 1035-1040.
- [56] Zaslavsky, Arkady, Charith Perera, and Dimitrios Georgakopoulos. "Sensing as a service and big data." arXiv preprint arXiv:1301.0159 (2013).
- [57] Bala, K., et al.. "Chat-bot for college management system using AI." International Research

Journal of Engineering and Technology 4.11 (2017): 2030-2033.

- [58] O'Leary, Daniel E. "Using AI in knowledge management: Knowledge bases and ontologies." IEEE Intelligent Systems and Their Applications 13.3 (1998): 34-39.
- [59] Lalwani, Tarun, et al.. "Implementation of a Chatbot System using AI and NLP." International Journal of Innovative Research in Computer Science & Technology (IJIRCST) Volume-6, Issue-3 (2018).
- [60] Mendes Rodrigues, Eduarda, and Natasa Milic-Frayling. "Socializing or knowledge sharing? Characterizing social intent in community question answering." Proceedings of the 18th ACM conference on Information and knowledge management. 2009.
- [61] Rajabion, Lila, et al.. "A new model for evaluating the effect of cloud computing on the e-learning development." Journal of Workplace Learning (2019).
- [62] Gjorgjevska, Emilija, and Georgina Mirceva. "Content Engineering for State-of-the-art SEO Digital Strategies by Using NLP and ML." 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA). IEEE, 2021.
- [63] Lee, Gichang, et al.. "Sentiment classification with word localization based on weakly supervised learning with a convolutional neural network." Knowledge-Based Systems 152 (2018): 70-82.
- [64] Lee, Joohong, Sangwoo Seo, and Yong Suk Choi. "Semantic relation classification via bidirectional lstm networks with entity-aware attention using latent entity typing." Symmetry 11.6 (2019): 785.
- [65] Wang, John, et al.. "Cloud Computing, Intelligent Business Process Management and Artificial Intelligence." international journal of data analysis and information systems 9 (2017): 1-12.
- [66] Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2021). Explainable ai: A review of machine learning interpretability methods. Entropy, 23(1), 18.